doi: 10.3389/fonc.2014.00007

Current challenges in the bioinformatics of single cell genomics

Luwen Ning[1], Geng Liu[2], Guibo Li[2], Yong Hou[2], Yin Tong[1] and Jiankui He[1]*

[1] *Department of Biology, South University of Science and Technology of China, Shenzhen, China*
[2] *BGI-Shenzhen, Shenzhen, China*

Edited by:
Tomer Kalisky, Bar-Ilan University, Israel

Reviewed by:
Myriam Alcalay, Istituto Europeo di Oncologia, Italy
Luisa Lanfrancone, European Institute of Oncology, Italy

****Correspondence:***
Jiankui He, Department of Biology, South University of Science and Technology of China, No 1088, Xueyuan Road, Xili, Nanshan District, Shenzhen, Guangdong 518055, China
e-mail: he.jk@sustc.edu.cn

Single cell genomics is a rapidly growing field with many new techniques emerging in the past few years. However, few bioinformatics tools specific for single cell genomics analysis are available. Single cell DNA/RNA sequencing data usually have low genome coverage and high amplification bias, which makes bioinformatics analysis challenging. Many current bioinformatics tools developed for bulk cell sequencing do not work well with single cell sequencing data. Here, we summarize current challenges in the bioinformatics analysis of single cell genomic DNA sequencing and single cell transcriptomes. These challenges include calling copy number variations, identifying mutated genes in tumor samples, reconstructing cell lineages, recovering low abundant transcripts, and improving the accuracy of quantitative analysis of transcripts. Development in single cell genomics bioinformatics analysis will promote the application of this technology to basic biology and medical research.

Keywords: single cell analysis, RNA sequencing, DNA sequencing, SNP, CNV

INTRODUCTION

Biologists have been interested in the heterogeneity between individual single cells at the molecular level in tissues and organs for a long time. For example, what is the difference between adjacent cells at the genetic and gene expression level in a tumor? What is different between cells at different developmental stages of human embryogenesis? These questions, and many similar questions, remain largely unanswered.

Novel sequencing technologies have rapidly advanced genomics studies in the past few years. Now, there are several exciting new techniques that enable us to sequence entire genomes at the single cell level. For example, multiple displacement amplification (MDA) has widely been used to amplify entire genomes from a few cells or even a single cell (1–3). Zong et al. recently described a multiple annealing, looping-based amplification cycle (MALBAC) method, which combines features of linear amplification methods with PCR (4). MALBAC has been shown to be capable of amplifying 93% of the genome of a single cell. Single cell sequencing technology has potentially broad applications in biology and medicine (5–7); for example, in the characterization of the earliest differentiation events in human embryogenesis (8); in the study of microorganisms that cannot be cultured (9–12); in transcriptome analysis of rare, circulating tumor cells (13–15); and in the study of tumor heterogeneity and microevolution (16–19).

The difference between single cell sequencing and bulk sequencing is that single cell sequencing needs an extra step that amplifies the genome from a single cell. It is this amplification process that makes the bioinformatics analysis of single cell sequencing data so challenging. The amplification process has two major technical problems. First, single cell amplification usually has a much lower genome coverage. Genomic regions that are not amplified will not be sequenced. Second, the amplification process will introduce artificial biases, with some genomic regions being amplified more than others. Because of these two reasons, many bioinformatics tools developed for bulk cell sequencing do not work well for single cell sequencing data. Nevertheless, as a revolutionary technology, single cell sequencing will quickly be applied in many biological and medical fields, and the bioinformatics community needs to act quickly to keep pace with the expected flood of single cell sequencing data.

In this review, we will describe the challenges in analyzing single cell DNA and RNA sequencing data. In addition, we will discuss the comparative analysis of multiple single cells.

SECTION I: BIOINFORMATICS IN SINGLE CELL DNA SEQUENCING

Single-nucleotide polymorphisms (SNPs) contribute most of the genetic variation to the human genome (20). SNPs associate with many monogenic and complex diseases, such as cancer, autoimmune disorders, diabetes, and Alzheimer's (21–24). Copy number variation (CNV) is another major type of genetic polymorphism (25) that has important roles in human health (26). CNV has been reported to be associated with various human diseases, such as tumors, autism, autoimmunity, systematic lupus erythematous, and other complex diseases (27–30). Analyzing DNA mutation and structural variation at the single cell level has been reported in a few recent studies (4, 31, 32). However, accurately calling SNP/CNV from single cell sequencing data remains challenging.

SNP CALLING IN SINGLE CELL DNA SEQUENCING

Calling SNPs in single cell data is a challenge that stems from the whole genome amplification (WGA) process itself. Typically, there are only about 6 pg of DNA in a single cell, and therefore, accurately measuring all of DNA information content from within

such a small amount is very difficult. WGA from single cells is an essential step during the library preparation for high throughput sequencing. Although the first WGA techniques appeared more than 10 years ago (33), current WGA methods still suffer from low coverage performance and errors during amplification, which present a hurdle in obtaining complete SNP information from a single cell.

LOW GENOME COVERAGE CAUSES SNP DROPOUT

Genome coverage is a measurement of the percentage of a genome covered by at least, one sequencing read. Current WGA techniques often have a lower genome coverage than bulk cell sequencing. For example, the MDA method can typically obtain an average genome coverage of 73% at a 25× sequencing depth (4). The recently developed MALBAC method enables significant improvements over traditional MDA methods, and can reach a genome coverage of 93% at a 30× sequencing depth (with an average genome coverage of 34% at a 25× sequencing depth) (4). The genome coverage of single cell amplification methods is still much lower than that of bulk cell sequencing, which can achieve more than 90% coverage at a 4× sequencing depth. SNPs in genomic regions that are not covered by sequencing reads will drop out of the analysis owing to these reasons. Furthermore, MDA methods suffer from a high ratio of allele dropout (ADO, alleles present in heterozygous samples not called out in the analysis). The ADO rate of MDA methods can be as high as 65%, as estimated by Zong et al. (4) Therefore, we should be very careful using single cell SNP calling results to perform further analyses such as those for gene ontology and pathway enrichment studies.

ERRORS IN AMPLIFICATION LEAD TO FALSE-POSITIVE SNP CALLING

Although WGA methods usually use a high fidelity polymerase enzyme, single cell sequencing still introduces a certain amount of false-positive error in calling SNPs. The MDA methods use Φ29 bacteriophage DNA polymerase, which has been shown to have a low error rate, approximately 10^{-5} per base (35). Zong et al. sequenced DNA from single cells amplified with both MALBAC and MDA methods, and found that the false-positive rate for genotyping single-nucleotide variants with MALBAC was about 40-fold higher than it was for MDA (36). They identified 2.2×10^6 SNPs in a single cell using GATK software (37), but the data contained 1.1×10^5 false positives, which means that one in 20 SNPs is artificial.

STRATEGIES FOR DEVELOPING SINGLE CELL SNP CALLING ALGORITHMS

Single-nucleotide polymorphisms calling algorithms for bulk cell samples have been studied extensively; among them, GATK (37), SNPdetector (38), SOAPsnp (39), and VarScan (40) are widely used. However, there is no SNP calling algorithm originally designed for single cell sequencing data. Researchers have used established software to call SNPs in a few recently published single cell studies. For example, Zong et al. (4) used GATK, while Xu et al. (19) used SOAPsnp to call SNPs in single cell sequencing samples. None of these methods, however, take the intrinsic properties of single cell amplification into consideration. To develop a SNP calling algorithm specifically designed for single cell sequencing, and to overcome the low genome coverage and high false-positive rate shortcomings, we have the following two suggestions: (1) the algorithm should be able to distinguish true SNPs from amplification errors; and (2) the algorithm should be able to call SNPs from low coverage sequencing.

CNV CALLING IN SINGLE CELL SEQUENCING

Copy number variation in genomes results in cells having an abnormal number of copies of one or more sections of DNA. Currently many software packages are available for calling CNVs in bulk DNA sequencing, such as CNV-seq (41), PenCNV (42), CNAseg (43), Readdepth (44), and cn.MOPS (45). However, few software packages and algorithms have been designed for single cell CNV calling, and the impact of amplification bias on CNV calling has not been systematically investigated.

AMPLIFICATION BIAS IN SINGLE CELL CNV CALLING

Multiple displacement amplification bias has been observed in several studies (46). For example, MDA has been reported to introduce hundreds of potentially confounding CNV artifacts that can obscure the detection of real variants (47). Many artifacts are reproducible, and may correlate with proximity to chromosome ends and GC content. The WGA4 (Sigma Genomeplex) method and MALBAC also have strong read fluctuations mapped to different genome regions. The majority of CNV calling methods in bulk cells are read count-based approaches. As an example, Readdepth is reported to be able to detect CNVs with sizes as small as 500 bp in 37× sequencing depths (44). However, in single cell sequencing the bin size must be increased to reduce read mapping bias when using read count-based approaches, owing to strong amplification biases. Navin et al. adopted a variable length bins method with a medium length of 54 kb (18). In MALBAC, the CNVs of single cell samples were analyzed at a resolution of 200 kb bin size (4). Baslan et al. proposed a protocol to analyze the CNV of single cell sequencing using a default bin size of 50 kb (48). The reproducibility of single cell CNV detection is also relatively low. Zong et al. (4) reported that the read number in 200 kb bins of two single cells from the same tissue has a correlation coefficient value less than 0.8.

STRATEGIES FOR ACCURATE SINGLE CELL CNV CALLING

To improve the quality and accuracy of calling CNV from single cell sequencing, we have the following suggestions: First, we need to carefully examine the bias generated in the genome amplification process. For example, in one study of the MDA method, recurrent MDA-induced copy number biases were reported to associate with sequence repeats and proximity to chromosome ends, increased GC content, and annotated CNVs (47). Once we know the pattern of artificial biases, we can develop algorithms to reduce the noise to call confident CNV assessment. Second, noise reduction problems also exist in other fields, such as signal processing and image processing, where noise reduction has been extensively studied. We can employ algorithms such as wavelet (49) and/or Fourier transformation (50) to single cell data to reduce noise. Third, pairwise comparisons of amplified products should help to reduce the number of artificial CNVs.

SECTION II: SINGLE CELL RNA SEQUENCING

The identity and function of a cell is determined by its entire RNA component. Ideally, transcriptome analysis should capture the exact quantity of all full-length RNAs of all classes, at single-base resolution, in an individual cell. Recently, several studies have reported single cell RNA sequencing analysis (51–55). We will discuss quantitative expression analysis, the detection of transcripts, and the identification of their splicing isoforms, using single-cell RNA sequencing in this section.

Several groups have demonstrated the application of single cell RNA sequencing in various biological systems (8, 51, 56–70) in recent years. For example, Tang et al. studied the blastomere cell using single cell RNA sequencing and found that 8–19% of the genes with multiple known transcript isoforms expressed at least two of those isoforms in the same blastomere cell (57). Ramsköld et al. applied Smart-Seq to study the gene expression profile of rare, circulating tumor cells from the blood of a melanoma patient, and found that the profile is highly correlated with those of melanoma cell lines, strongly indicating that the circulating cells originated from a melanoma tumor (15). Shalek et al. studied the transcriptomics of single immune cells and revealed a bimodality in expression and splicing (64).

The computational measurement of quantitative gene expression has been extensively studied in bulk cell sequencing analysis. Gene expression can be calculated from the number of sequencing reads mapped to a particular gene region. Current approaches in the analysis of quantitative gene expression include two steps: mapping RNA sequencing reads to gene regions, and calculating expression levels. In the first step, many tools have been developed for sequencing read mapping, such as TopHat (34), RUM (71), SpliceMap (72), MapSplice (73), GSNAP (74), BLAT (75), Bowtie (76), SOAP (77), and BWA (78). In the second step, RPKM (reads per kilobase per million reads) (79) and FPKM (fragments per kilobase per million fragments) (80) are commonly used to measure gene expression levels. Yan et al. used BWA to align reads to a reference genome (70) in a recent single cell RNA sequencing project. They selected genes with RPKM ≥ 0.1 for further analysis. Picelly et al. adopted STAR (81) to align sequence reads in another recent project, and then used RPKM for genes (82) to calculate RPKM. Shalek et al. used Tophat1 (34) to map reads to a reference genome, then used RSEM (83) to obtain the expression level of TPM (transcripts per million), and then used MISO (84) to locate the alternate splicing events (61).

The challenge of single cell RNA bioinformatics analysis is mainly due to the bias and distortion in the whole transcript amplification process. There are three major issues in single, whole transcript amplification: (1) amplification cannot generate full-length cDNAs; (2) transcripts are not amplified at the same ratio; and, (3) low abundant transcripts are difficult to detect.

It is usually difficult to get full-length transcripts in single cell RNA sequencing, and, therefore, 3′-end biases are often generated. Tang et al. only obtained RNA transcripts shorter than 3 kb in single mouse blastomere cells, missing 36% of the expressed genes (51). The median read coverage across expressed transcripts is 53.8% in the Quartz-Seq method, compared with 84.4% in conventional RNA sequencing (60). Ramsköld et al. demonstrated that the Smart-Seq technique can identify 40% of all full-length transcripts (15). The FPKM/RPKM values, which are commonly used to measure gene expression levels, do not consider bias across the transcripts, and therefore, may not suitable for single cell RNA sequencing in bioinformatics analysis. Furthermore, the pronounced 3′-end bias of whole transcript amplification may hamper the ability to identify alternate splicing differences in single cells.

GC content and cDNA length distribution may also induce artificial biases during whole transcript amplification. For example, Sasagawa et al. showed that unamplified isoforms from Quartz-Seq have a higher GC content, with value of 52.1%, versus the mean GC content of the amplified isoforms, with a value of 50.2%, which indicates that high GC content RNAs are difficult to amplify (60). They also found that amplified cDNAs have a longer length than the unamplified isoforms that correspond to the cDNA. Because of these reasons, we need to validate to what extent single-cell transcriptomes faithfully represent the RNA populations they reflect before amplification.

Bulk cell samples are usually sequenced at high depth to obtain low-abundance transcripts. However, in single cell sequencing, the low-abundance transcripts may not be so easily amplified. For example, the transcription detection rate is about 80% in Quartz-Seq and about 55% in Smart-Seq (60). Picelili found the observed variability between cells was mainly of a technical nature for low-abundance transcripts, whereas in medium- and high-abundance transcripts, variability between cells was mainly biological (58).

New methods are needed to generate an unbiased quantitative measure of transcript expression in single-cell transcriptomics analysis. Recently several new amplification protocols have been developed, but bias still exists to a certain extent. Based on our own experience, we suggest the following two potential solutions, which may help to address the problems of amplification bias and low coverage:

1. Consider a new standard expression level measurement beyond RPKM/FPKM in single cell RNA sequencing. Single cell RNA sequencing data usually has 3′ and/or 5′ biases; therefore, measuring expression levels using full-length transcripts may be inappropriate. One possible solution is to normalize expression levels by coverage lengths instead of by using full-length transcripts.
2. Reduce amplification biases by developing new bioinformatics approaches. We can systematically investigate how bias is generated during amplification to discover the patterns of bias. Machine learning (85–92) may be a powerful tool to study the distribution of bias and to predict amplification bias.

SECTION III: COMPARATIVE ANALYSIS OF SINGLE CELLS

One important goal of single cell technology is to discover heterogeneity among cells (69, 93, 94). Dozens of single cells are usually sequenced in a single cell genomics project. The heterogeneity between cells can thus be found by comparative analysis between the single cells employed. Here we will describe the comparative analysis of single cells in the development and lineage structure of tumor and early embryonic development.

DEVELOPMENT AND LINEAGE STRUCTURE OF TUMORS

Genetic heterogeneity is very common in tumors, and is important information for reconstructing evolutionary history. This information may be averaged out in bulk cell sequencing (95). However, the comparative analysis of sequencing data from multiple single cells is a much more powerful technique for studying tumor population structure and evolution (18).

Navin et al. applied single-nucleus sequencing to investigate tumor population structure and evolution in two human breast cancer cases, and found that tumors grow by punctuated clonal expansions with few persistent intermediates (18). Hou et al. inferred tumor monoclonal origin in an essential thrombocythemia patient using single cell exome sequencing (96).

Although single cell sequencing can provide robust information regarding tumor heterogeneity and evolution, there are still several technical problems to resolve. The accuracy and sensitivity of detecting single cell variation can significantly affect single cell population analysis. In general, only mutations observed in multiple cells can confidently be considered real mutations. As a result, some rare mutations or cell clones may not be able to be identified with a reasonable confidence level. Another question is how many cells should be sequenced in a single cell sequencing project? The sequencing cost of a single cell is still relatively high, and we need a statistic model to evaluate the appropriate number of single cells in a project.

EMBRYONIC DEVELOPMENT

Single cell RNA sequencing can provide insight into the dynamic expression of key genes to explore the relationship among the different stages of stem cells (57). Tang et al. traced the derivation of embryonic stem cells from the inner cell mass by single-cell RNA sequencing analysis (97), providing insight into the dynamic molecular changes that accompany cell fate changes based on the expression of both mRNA and microRNA. Xue et al. reported a comprehensive analysis of transcriptome dynamics from oocyte to morula in both human and mouse embryos, and identified embryonic genome activation events (8). These results provide valuable resources for dissecting gene regulatory mechanisms, and for understanding the underlying progressive development of early mammalian embryos. For example, single cell RNA sequencing has a great potential for discovering previously unrecognized biological distinctions between two-cell, four-cell, eight-cell, and later stages of embryogenesis. However, the high level of noise inherent in single cell genomics is hard to address, because of technical limitations in both experimental preparations and computational approaches, due to biological reasons and the limited amount of input material available. Future studies based on hundreds or thousands of single cells with new bioinformatics approaches will enable analyses to reconstruct intracellular genetic circuits, enumerate and redefine cell developmental states and types, and understand cellular decision-making on a genomic scale.

CONCLUSION

Single cell genomics analysis not only provides a more precise measurement, but is also a decisive move toward a fundamental understanding of the biology of cells. The ever-increasing power of DNA sequencing technology means that it will soon be possible to sequence every nucleic acid in many thousands of cells.

At present, single cell sequencing techniques still have two major shortcomings: low genome coverage, and high amplification bias. Despite the limitations, these still-evolving technologies will eventually revolutionize research in oncology, neuroscience, cell development, and microbiology. Partly through innovations in microfluidics (98, 99) and next generation sequencing technologies, we expect that the primary nucleic acid sequence analysis of single cell genomic DNAs and RNAs will be solved in a few years.

However, few existing bioinformatics software packages exist for the purpose of single cell genomics data analysis. Continued advances in the application of singe cell sequencing technologies in biological research will require development of new algorithms and software able to handle the specific characteristics of these technologies. In particular, we need tools to evaluate the performance of different single cell sequencing technologies. Technical standards should be built to evaluate the genome coverage and amplification biases, so that the results from different technologies can be compared with each other. Meanwhile, new tools are needed to manage the large amounts of data generated by single cell sequencing technologies, which is expected to be one order magnitude larger than regular bulk sequencing projects. An open-source and shared model will accelerate the progress by allowing the scientific community to join forces in addressing the challenges and promises of the new technologies.

ACKNOWLEDGMENTS

The project is supported by the National Natural Science Foundation of China (No. 31200688).

REFERENCES

1. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* (2002) **99**(8):5261–6. doi:10.1073/pnas.082089499
2. Spits C, Le Caignec C, De Rycke M, Van Haute L, Van Steirteghem A, Liebaers I, et al. Whole-genome multiple displacement amplification from single cells. *Nat Protoc* (2006) **1**(4):1965–70. doi:10.1038/nprot.2006.326
3. Lasken RS. Single-cell genomic sequencing using multiple displacement amplification. *Curr Opin Microbiol* (2007) **10**(5):510–6. doi:10.1016/j.mib.2007.08.005
4. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* (2012) **338**(6114):1622–6. doi:10.1126/science.1229164
5. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* (2013) **14**(9):618–30. doi:10.1038/nrg3542
6. Speicher MR. Single-cell analysis: toward the clinic. *Genome Med* (2013) **5**(8):1–3. doi:10.1186/gm478
7. Bermudez MG, Piyamongkol W, Tomaz S, Dudman E, Sherlock JK, Wells D. Single-cell sequencing and mini-sequencing for preimplantation genetic diagnosis. *Prenat Diagn* (2003) **23**(8):669–77. doi:10.1002/pd.658
8. Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* (2013) **500**(7464):593–7. doi:10.1038/nature12364
9. Blainey PC. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev* (2013) **37**(3):407–27. doi:10.1111/1574-6976.12015
10. Dodsworth JA, Blainey PC, Murugapiran SK, Swingley WD, Ross CA, Tringe SG, et al. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun* (2013) **4**:1854. doi:10.1038/ncomms2884

11. Mason OU, Hazen TC, Borglin S, Chain PS, Dubinsky EA, Fortney JL, et al. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J* (2012) **6**(9):1715–27. doi:10.1038/ismej.2012.59
12. Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS. Genomic sequencing of single microbial cells from environmental samples. *Curr Opin Microbiol* (2008) **11**(3):198–204. doi:10.1016/j.mib.2008.05.006
13. Heitzer E, Auer M, Gasch C, Pichler M, Ulz P, Hoffmann EM, et al. Complex tumor genomes inferred from single circulating tumor cells by array-CGH and next-generation sequencing. *Cancer Res* (2013) **73**(10):2965–75. doi:10.1158/0008-5472.CAN-12-4140
14. Bidard F-C, Weigelt B, Reis-Filho JS. Going with the flow: from circulating tumor cells to DNA. *Sci Transl Med* (2013) **5**(207):s14–14. doi:10.1126/scitranslmed.3006305
15. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* (2012) **30**(8):777–82. doi:10.1038/nbt.2282
16. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* (2011) **29**(12):1120–7. doi:10.1038/nbt.2038
17. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer* (2012) **12**(5):323–34. doi:10.1038/nrc3261
18. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. *Nature* (2011) **472**(7341):90–4. doi:10.1038/nature09807
19. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* (2012) **148**(5):886–95. doi:10.1016/j.cell.2012.02.025
20. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* (2012) **491**:1. doi:10.1038/nature11632
21. Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari A, Riley J, et al. SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around *APOE* in Alzheimer disease. *Am J Hum Genet* (2000) **67**(2):383–94. doi:10.1086/303003
22. Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (*PTPN22*) is associated with rheumatoid arthritis. *Am J Hum Genet* (2004) **75**(2):330–7. doi:10.1086/422827
23. Saxena R, Gianniny L, Burtt NP, Lyssenko V, Giuducci C, Sjögren M, et al. Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. *Diabetes* (2006) **55**(10):2890–5. doi:10.2337/db06-0381
24. Zhu Y, Spitz MR, Lei L, Mills GB, Wu X. A single nucleotide polymorphism in the matrix metalloproteinase-1 promoter enhances lung cancer susceptibility. *Cancer Res* (2001) **61**(21):7825–9.
25. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature* (2006) **444**(7118):444–54. doi:10.1038/nature05329
26. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* (2009) **10**:451–81. doi:10.1146/annurev.genom.9.081307.164217
27. Frank B, Bermejo JL, Hemminki K, Sutter C, Wappenschmidt B, Meindl A, et al. Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. *Carcinogenesis* (2007) **28**(7):1442–5. doi:10.1093/carcin/bgm033
28. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* (2009) **459**(7246):569–73. doi:10.1038/nature07953
29. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet* (2007) **80**(6):1037–54. doi:10.1086/518257
30. Schaschl H, Aitman T, Vyse T. Copy number variation in the human genome and its implication in autoimmunity. *Clin Exp Immunol* (2009) **156**(1):12–6. doi:10.1111/j.1365-2249.2008.03865.x
31. Lorthongpanich C, Cheow LF, Balu S, Quake SR, Knowles BB, Burkholder WF, et al. Single-cell DNA-methylation analysis reveals epigenetic chimerism in preimplantation embryos. *Science* (2013) **341**(6150):1110–2. doi:10.1126/science.1240617
32. Lu S, Zong C, Fan W, Yang M, Li J, Chapman AR, et al. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* (2012) **338**(6114):1627–30. doi:10.1126/science.1229112
33. Zhang L, Cui X, Schmitt K, Hubert R, Navidi W, Arnheim N. Whole genome amplification from a single cell: implications for genetic analysis. *Proc Natl Acad Sci U S A* (1992) **89**(13):5847–51. doi:10.1073/pnas.89.13.5847
34. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* (2009) **25**:1105–11. doi:10.1093/bioinformatics/btp120
35. Esteban J, Salas M, Blanco L. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J Biol Chem* (1993) **268**(4):2719–26.
36. Lasken RS. Single-cell sequencing in its prime. *Nat Biotechnol* (2013) **31**(3):211–2. doi:10.1038/nbt.2523
37. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* (2010) **20**(9):1297–303. doi:10.1101/gr.107524.110
38. Zhang J, Wheeler DA, Yakub I, Wei S, Sood R, Rowe W, et al. SNPdetector: a software tool for sensitive and accurate SNP detection. *PLoS Comput Biol* (2005) **1**(5):e53. doi:10.1371/journal.pcbi.0010053
39. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res* (2009) **19**(6):1124–32. doi:10.1101/gr.088013.108
40. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* (2009) **25**(17):2283–5. doi:10.1093/bioinformatics/btp373
41. Xie C, Tammi M. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* (2009) **10**(1):80. doi:10.1186/1471-2105-10-80
42. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* (2007) **17**(11):1665–74. doi:10.1101/gr.6861907
43. Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, Tavaré S. CNAseg – a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* (2010) **26**(24):3051–8. doi:10.1093/bioinformatics/btq587
44. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* (2011) **6**(1):e16327. doi:10.1371/journal.pone.0016327
45. Klambauer G, Schwarzbauer K, Mayr A, Clevert D-A, Mitterecker A, Bodenhofer U, et al. MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res* (2012) **40**(9):e69–e. doi:10.1093/nar/gks003
46. Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, et al. Unbiased whole-genome amplification directly from clinical samples. *Genome Res* (2003) **13**(5):954–64. doi:10.1101/gr.816903
47. Pugh T, Delaney A, Farnoud N, Flibotte S, Griffith M, Li H, et al. Impact of whole genome amplification on analysis of copy number variants. *Nucleic Acids Res* (2008) **36**(13):e80–e. doi:10.1093/nar/gkn378
48. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, et al. Genome-wide copy number analysis of single cells. *Nat Protoc* (2012) **7**(6):1024–41. doi:10.1038/nprot.2012.039
49. Xu Y, Weaver JB, Healy DM, Lu J. Wavelet transform domain filters: a spatially selective noise filtration technique. *Image Proc IEEE Trans* (1994) **3**(6):747–58. doi:10.1109/83.336245
50. Boll S. Suppression of acoustic noise in speech using spectral subtraction. *Acoust Speech Signal Proc IEEE Trans* (1979) **27**(2):113–20. doi:10.1109/TASSP.1979.1163209
51. Tang F, Barbacioru C, Nordman E, Li B, Xu N, Bashkirov VI, et al. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* (2010) **5**(3):516–35. doi:10.1038/nprot.2009.236

52. Crobu F, Latini V, Marongiu MF, Sogos V, Scintu F, Porcu S, et al. Differentiation of single cell derived human mesenchymal stem cells into cells with a neuronal phenotype: RNA and microRNA expression profile. *Mol Biol Rep* (2012) **39**(4):3995–4007. doi:10.1007/s11033-011-1180-9
53. Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol* (2013) **14**(1):R7. doi:10.1186/gb-2013-14-1-r7
54. Subkhankulova T, Gilchrist MJ, Livesey FJ. Modelling and measuring single cell RNA expression levels find considerable transcriptional differences among phenotypically identical cells. *BMC Genomics* (2008) **9**:268. doi:10.1186/1471-2164-9-268
55. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* (2014) **9**(1):171–81. doi:10.1038/nprot.2014.006
56. Hashimshony T, Wagner F, Sher N, Yanai ICEL-. Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports* (2012) **2**(3):666–73. doi:10.1016/j.celrep.2012.08.003
57. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* (2009) **6**(5):377–82. doi:10.1038/nmeth.1315
58. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* (2013) **10**:1096–8. doi:10.1038/nmeth.2639
59. Pan X, Durrett RE, Zhu H, Tanaka Y, Li Y, Zi X, et al. Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc Natl Acad Sci U S A* (2013) **110**(2):594–9. doi:10.1073/pnas.1217322109
60. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. *Genome Biol* (2013) **14**(4):R31. doi:10.1186/gb-2013-14-4-r31
61. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* (2013) **498**(7453):236–40. doi:10.1038/nature12172
62. Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, Sexton DW, et al. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol* (2013) **31**(8):748–52. doi:10.1038/nbt.2642
63. Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, et al. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* (2013) **11**(1):41–6. doi:10.1038/nmeth.2694
64. Voet T, Kumar P, Van Loo P, Cooke SL, Marshall J, Lin M-L, et al. Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Res* (2013) **41**(12):6119–38. doi:10.1093/nar/gkt345
65. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* (2013) **10**(11):1093–5. doi:10.1038/nmeth.2645
66. Hebenstreit D. Methods, challenges and potentials of single cell RNA-seq. *Biology* (2012) **1**(3):658–67. doi:10.3390/biology1030658
67. Spaethling JM, Eberwine JH. Single-cell transcriptomics for drug target discovery. *Curr Opin Pharmacol* (2013) **13**(5):786–90. doi:10.1016/j.coph.2013.04.011
68. Katayama S, Töhönen V, Linnarsson S, Kere J. SAMstrt: Statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* (2013) **29**(22):2943–5. doi:10.1093/bioinformatics/btt511
69. Ren S-C, Qu M, Sun Y-H. Investigating intratumour heterogeneity by single-cell sequencing. *Asian J Androl* (2013) **15**(6):729–34. doi:10.1038/aja.2013.106
70. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* (2013) **20**(9):1131–9. doi:10.1038/nsmb.2660
71. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* (2011) **27**(18):2518–28. doi:10.1093/bioinformatics/btr427
72. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* (2010) **38**(14):4570–8. doi:10.1093/nar/gkq211
73. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* (2010) **38**(18):e178. doi:10.1093/nar/gkq622
74. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* (2010) **26**(7):873–81. doi:10.1093/bioinformatics/btq057
75. Kent WJ. BLAT – the BLAST-like alignment tool. *Genome Res* (2002) **12**(4):656–64. doi:10.1101/gr.229202
76. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* (2012) **9**(4):357–9. doi:10.1038/nmeth.1923
77. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics* (2008) **24**(5):713–4. doi:10.1093/bioinformatics/btn025
78. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009) **25**(14):1754–60. doi:10.1093/bioinformatics/btp324
79. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* (2008) **5**(7):621–8. doi:10.1038/nmeth.1226
80. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* (2010) **28**(5):511–5. doi:10.1038/nbt.1621
81. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* (2013) **29**(1):15–21. doi:10.1093/bioinformatics/bts635
82. Ramsköld D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* (2009) **5**(12):e1000598. doi:10.1371/journal.pcbi.1000598
83. Li B, Dewey C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* (2011) **12**(1):323. doi:10.1186/1471-2105-12-323
84. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* (2010) **7**(12):1009–15. doi:10.1038/nmeth.1528
85. Bhaskar H, Hoyle DC, Singh S. Machine learning in bioinformatics: a brief survey and recommendations for practitioners. *Comput Biol Med* (2006) **36**(10):1104–25. doi:10.1016/j.compbiomed.2005.09.002
86. Hamelryck T. Probabilistic models and machine learning in structural bioinformatics. *Stat Methods Med Res* (2009) **18**(5):505–26. doi:10.1177/0962280208099492
87. Inza I, Calvo B, Armananzas R, Bengoetxea E, Larranaga P, Lozano JA. Machine learning: an indispensable tool in bioinformatics. *Methods Mol Biol* (2010) **593**:25–48. doi:10.1007/978-1-60327-194-3_2
88. Kurgan L, Zhou Y. Machine learning models in protein bioinformatics. *Curr Protein Pept Sci* (2011) **12**(6):455. doi:10.2174/138920311796957621
89. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. *Brief Bioinform* (2006) **7**(1):86–112. doi:10.1093/bib/bbk007
90. Mattison HA, Stewart T, Zhang J. Applying bioinformatics to proteomics: is machine learning the answer to biomarker discovery for PD and MSA? *Mov Disord* (2012) **27**(13):1595–7. doi:10.1002/mds.25189
91. Peterson LE, Chen XW. Machine learning in biomedicine and bioinformatics. *Int J Data Min Bioinform* (2009) **3**(4):363–4.
92. Valentini G, Tagliaferri R, Masulli F. Computational intelligence and machine learning in bioinformatics. *Artif Intell Med* (2009) **45**(2–3):91–6. doi:10.1016/j.artmed.2008.08.014
93. Wilson JL, Suri S, Singh A, Rivet CA, Lu H, McDevitt TC. Single-cell analysis of embryoid body heterogeneity using microfluidic trapping array. *Biomed Microdevices* (2013). doi:10.1007/s10544-013-9807-3
94. Abdallah BY, Horne SD, Stevens JB, Liu G, Ying AY, Vanderhyden B, et al. Single cell heterogeneity: why unstable genomes are incompatible with average profiles. *Cell Cycle* (2013) **12**(23). doi:10.4161/cc.26580
95. Park SY, Gonen M, Kim HJ, Michor F, Polyak K. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J Clin Invest* (2010) **120**(2):636–44. doi:10.1172/JCI40724
96. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* (2012) **148**(5):873–85. doi:10.1016/j.cell.2012.02.028
97. Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, et al. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* (2010) **6**(5):468–78. doi:10.1016/j.stem.2010.03.015

98. Wang J, Fan HC, Behr B, Quake SR. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* (2012) **150**(2):402–12. doi:10.1016/j.cell.2012.06.030
99. Gole J, Gore A, Richards A, Chiu Y-J, Fung H-L, Bushman D, et al. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat Biotechnol* (2013) **31**(12):1126–32. doi:10.1038/nbt.2720

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 October 2013; paper pending published: 22 December 2013; accepted: 12 January 2014; published online: 27 January 2014.
*Citation: Ning L, Liu G, Li G, Hou Y, Tong Y and He J (2014) Current challenges in the bioinformatics of single cell genomics. Front. Oncol. **4**:7. doi: 10.3389/fonc.2014.00007 This article was submitted to Molecular and Cellular Oncology, a section of the journal Frontiers in Oncology.*